**BIS2216 Data Mining and Knowledge Discovery Fundamentals**
**Semester August 2020**
**Coursework (15% of Total Assessment)**

**Name: Chua Wen Soong ID: 18032573**

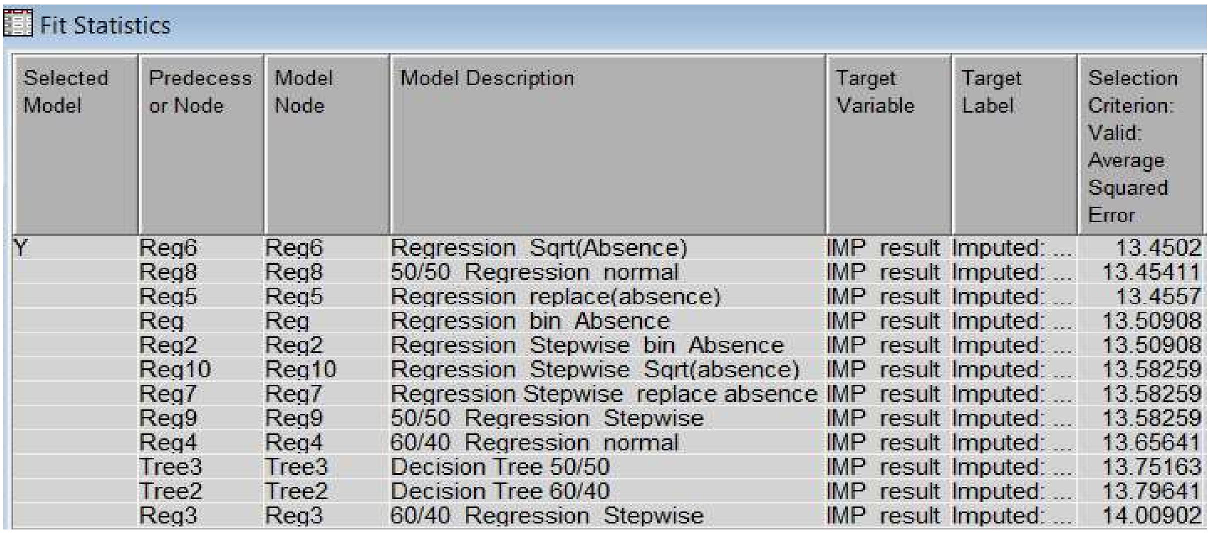
## 1. Data Preparation

| Attribute Processed | Rationale for Processing | Methods for Processing |
|---|---|---|
| Age | 1 observation has an age value of '1'. It is most likely incorrectly imputed. | The value of '1' was changed to the mean of age. |
| Gender | 1 observation has a value of "female" while the rest is simply "F" or "M". | The value "female" was changed to "F". |
| Results | 1 observation has a missing value for results. | The missing value was replaced by imputing the mean of results. |

## 2. Data Modelling

| No. | Modelling Technique | Partition Ratio | Other preparation methods applied | Model performance error |
|---|---|---|---|---|
| 1 | Regression | 60/40 | None | 13.65641 |
| 2 | Regression | 60/40 | Stepwise model selection | 14.00902 |
| 3 | Decision Tree | 60/40 | None | 13.79641 |
| 4 | Regression | 50/50 | None | 13.45411 |
| 5 | Regression<br>Stepwise model selection | 50/50 | None | 13.58259 |
| 6 | Decision Tree | 50/50 | None | 13.75163 |
| 7 | Regression | 50/50 | Absence was binned into:<br>1: absences <0 or missing<br>2: 0<= absences < 2<br>3: 2<= absences < 6<br>4: absences >= 6 | 13.50908 |
| 8 | Regression<br>Stepwise model selection | 50/50 | Absence was binned into:<br>1: absences <0 or missing<br>2: 0<= absences < 2<br>3: 2<= absences < 6<br>4: absences >= 6 | 13.50908 |
| 9 | Regression | 50/50 | Absence with values 70 or more are replaced with computed | 13.4557 |
| 10 | Regression<br>Stepwise model selection | 50/50 | Absence with values 70 or more are replaced with computed | 13.58259 |
| 11 | Regression | 50/50 | Absence was transformed with formula sqrt(Absence) | 13.4502 |
| 12 | Regression<br>Stepwise model selection | 50/50 | Absence was transformed with formula sqrt(Absence) | 13.58259 |

### 3. Model Comparison



Fit Statistics

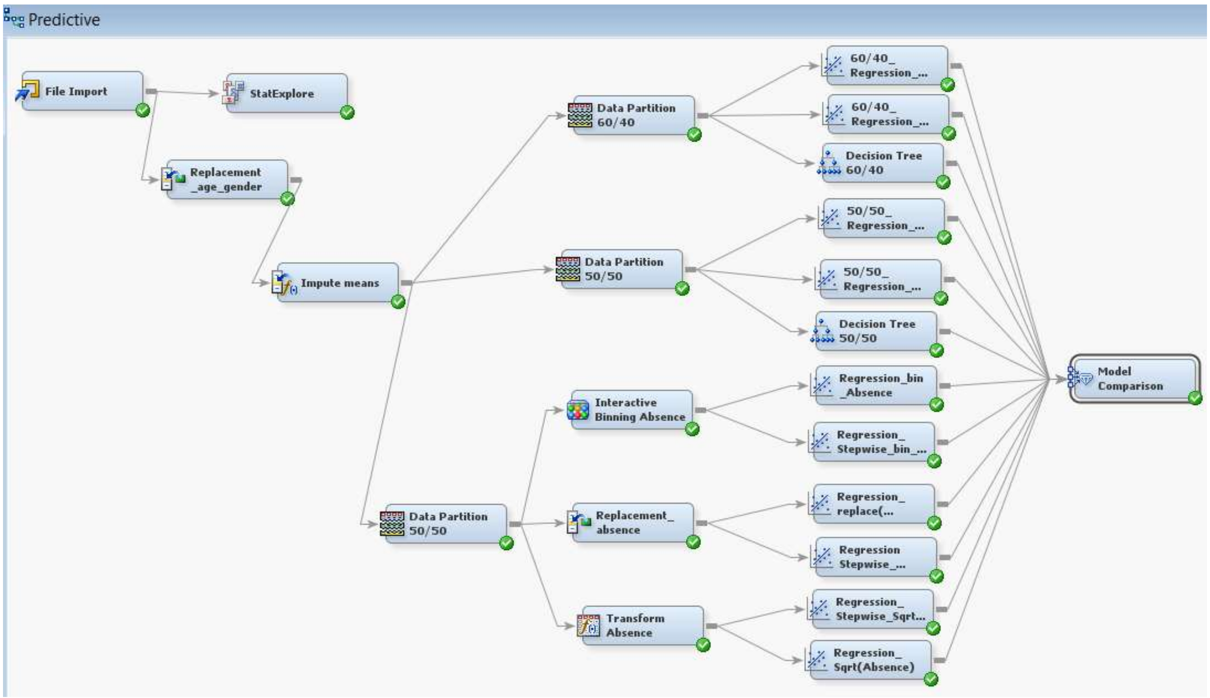
| Selected Model | Predecessor Node | Model Node | Model Description | Target Variable | Target Label | Selection Criterion: Valid: Average Squared Error |
|---|---|---|---|---|---|---|
| Y | Reg6 | Reg6 | Regression Sqrt(Absence) | IMP result | Imputed: ... | 13.4502 |
|  | Reg8 | Reg8 | 50/50 Regression normal | IMP result | Imputed: ... | 13.45411 |
|  | Reg5 | Reg5 | Regression replace(absence) | IMP result | Imputed: ... | 13.4557 |
|  | Reg | Reg | Regression bin Absence | IMP result | Imputed: ... | 13.50908 |
|  | Reg2 | Reg2 | Regression Stepwise bin Absence | IMP result | Imputed: ... | 13.50908 |
|  | Reg10 | Reg10 | Regression Stepwise Sqrt(absence) | IMP result | Imputed: ... | 13.58259 |
|  | Reg7 | Reg7 | Regression Stepwise replace absence | IMP result | Imputed: ... | 13.58259 |
|  | Reg9 | Reg9 | 50/50 Regression Stepwise | IMP result | Imputed: ... | 13.58259 |
|  | Reg4 | Reg4 | 60/40 Regression normal | IMP result | Imputed: ... | 13.65641 |
|  | Tree3 | Tree3 | Decision Tree 50/50 | IMP result | Imputed: ... | 13.75163 |
|  | Tree2 | Tree2 | Decision Tree 60/40 | IMP result | Imputed: ... | 13.79641 |
|  | Reg3 | Reg3 | 60/40 Regression Stepwise | IMP result | Imputed: ... | 14.00902 |

### 4 and 5. Best Model Presentation and Interpretation

The best model is a standard linear regression with a square rooted absence variable. This model has an average squared error of 13.4502. In the analysis of variance, it has a p-value of <0.0001, which means it is statistically significant. Its adjusted r-squared value is 0.0879 which explains 8.79% of the total variation of the model. With the significance level p > 0.05, the analysis of Maximum Likelihood Estimates produces a linear regression line:

y = 17.6329 – 0.3678(age) + 1.1861(Mjob_health) – 0.0469(goout) + 0.8237(studytime)

### 6. Screenshot of the whole modelling process


Predictive
File Import
StatExplore
Replacement _age_gender
Impute means
Data Partition 60/40
Data Partition 50/50
Data Partition 50/50
Interactive Binning Absence
Replacement_ absence
Transform Absence
60/40_ Regression_...
60/40_ Regression_...
Decision Tree 60/40
50/50_ Regression_...
50/50_ Regression_...
Decision Tree 50/50
Regression_bin _Absence
Regression_ Stepwise_bin_...
Regression_ replace(...
Regression Stepwise_...
Regression_ Stepwise_Sqrt...
Regression_ Sqrt(Absence)
Model Comparison